A Comparison of Simultaneous Confidence Intervals to Identify Handwritten Digits

Nicolle Clements, Department of Decision System Sciences, Saint Joseph’s University, Philadelphia, PA, USA

ABSTRACT

This paper evaluates the use of several known simultaneous confidence interval methods for the automated recognition of handwritten digits from data in a well-known handwriting database. Contained in this database are handwritten digits, 0 through 9, that were obtained from 42,000 participants’ writing samples. The objective of the analyses is to utilize statistical testing procedures that can be easily automated by a computer to recognize which digit was written by a subject. The methodologies discussed in this paper are designed to be sensitive to Type I errors and will control an overall measure of these errors, called the Familywise Error Rate. The procedures were constructed based off of a training portion of the data set, then applied and validated on the remaining testing portion of the data.

Keywords: Digit Recognition, Familywise Error Rate, Multiple Testing, Simultaneous Confidence Intervals, Type I Errors

1. INTRODUCTION

Automatic handwriting recognition is the technique by which a computer system can recognize characters and other symbols written by hand in one’s natural handwriting. The role of automatic handwriting recognition, of both alphabetic characters and numeric digits, is increasingly important as today’s technologies continue to improve. There are an enormous amount of applications of handwriting recognition, including the automatic scanning of personal checks at an ATM to be deposited into a bank account. Other applications include handwriting recognition on devices such as PDA’s and tablet PC’s where a stylus-pen is used to write on a screen, after which the computer turns the handwriting into digital text.

Another noteworthy application of handwriting recognition is signature verification. This is important because every year, millions of dollars are lost to fraudulent credit card charges, which could be prevented by more stringent signature verification policies. For example, many store clerks do not routinely check the signature of a customer against that of his/her credit card. Even if signature verification was regularly conducted, the clerk’s knowledge of

DOI: 10.4018/ijbir.2014070103
handwriting forgery would probably be limited, and thus the verification would be superficial. Signature verification, if done by specialized computer software, could do a much better analysis of the signature than any human specialist could ever do and might lessen the burden on the criminal justice system, which frequently investigates accusations of signature forgery (Huber & Headrick, 1999).

A few statistical techniques have been proposed within the handwriting recognition community, such as clustering procedures with Hidden Markov Models, Neural Network Models (Morasso, et al., 1993), maximum likelihood estimators (Sas & Kurzynski, 2007), and feature extraction methods using distance measures such as Kullback-Liebler. Several previously studied statistical handwriting identification models involve hierarchal clustering techniques. Nosary, et al. (2003) proposed a probabilistic approach to define clusters. In this study, each handwritten character or digit uses an approach to learn the probabilities that a character belongs to a given cluster.

Another statistical clustering approach was developed in Smyth (1997), where an algorithm was presented to cluster sequences into a predefined number of clusters, along with a preliminary method to find the numbers of clusters through cross-validation using a Monte Carlo estimation. This theoretical approach relies on iterative re-estimation of parameters via an instance of the expectation–maximization (EM) algorithm, which requires careful initialization. Furthermore, the structure of the model is limited to a mixture model of fixed-length left-right Hidden Markov Models, which may not correctly model sequences of varying length in the data. The idea of using Hidden Markov Models for clustering handwritten characters was later tackled by Perrone & Connell (2000), but their approach also depends on initialization parameters, thus some supervised information is needed to achieve good performance.

A research group at George Mason University and Gannon Technologies, under funding from the FBI, developed the system known as FLASH ID, which stands for Forensic Language-independent Analysis System for Handwriting Identification (Saunders, et al., 2011). The method consists of extracting features from graphs of characters and digits, building a graph feature vector, and identifying the unknown character or digit graph by matching it against a database containing a set of known character/digit graphs. These graphs are denoted as isocodes, which are built using nodes as the ends and cross-points of curves and the curves as the edges. The distribution of the data sample of isocodes is then compared to the population distribution using the Kullback-Liebler distance.

Many of these aforementioned techniques were successfully developed with the priority to have high accuracy rate. However, the goal of this paper is to introduce handwriting recognition techniques, applied to the 10 numeric digits, which focus on the error rate, instead of the accuracy rate. In many applications it is necessary to have a tight control on the error rate; for example, the criminal justice system’s investigation of handwriting forgery when accusing a suspect of fraudulent charges. The United States’ criminal justice system is based on the notion “innocent until proven guilty,” so any handwriting recognition software used in this application must be very cautious when making errors, or, in other words, cautious when identifying a handwriting sample as fraudulent (guilty), when in fact it is not (innocent).

There has already been some debate of how an individual’s type of writing plays a role in the United States Court System. Some United States federal district court judges feel there is a lack proficiency in identifying writers of hand-written documents. In the case of the Unites States vs. Jeffrey H. Feingold, error rates in Kam & Lin (2003) on analyzing handwriting were specifically questioned regarding their applicability. In that case, the court called a break in the trial, lasting several months, so that the results could be reexamined. This case demonstrates the need for an easily automated handwriting recognition procedure that still has a tight control on errors, so that such long breaks in a trial are unnecessary.
Predictive Skill Based Call Routing Using Multi-Label Classification Techniques
[www.igi-global.com/article/predictive-skill-based-call-routing-using-multi-label-classification-techniques/197404?camid=4v1a](www.igi-global.com/article/predictive-skill-based-call-routing-using-multi-label-classification-techniques/197404?camid=4v1a)

Visualization of High-Level Associations from Twitter Data
[www.igi-global.com/chapter/visualization-of-high-level-associations-from-twitter-data/142646?camid=4v1a](www.igi-global.com/chapter/visualization-of-high-level-associations-from-twitter-data/142646?camid=4v1a)